Problem sets for finals: Learn how things are connected in the ~~ML life~~-cycle

Dr. Ali Arsanjani

Applied ML

12/8/20

1. Smooth Sailing AV

- Remove redundant imports and connections to various googledrives
- So we do not have to reconnect 50 times!!

```
from toxicityclass1 import GirlsWhoCode_Toxicity
from Topics_with_LDA_Bigram import Topics_with_LDA_Bigram
#from girlswhocode_bias import Gwc_Bias
from bias import Gwc_Bias
from girlswhocode_political_affiliation import Girlswhocode_PoliticalAfiiliation
```

```
Mounted at /content/drive
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

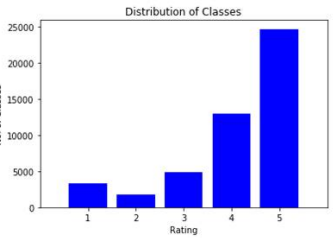
1. Smooth Sailing AV

- Make it uniform,
- Git or not git!

- ML + SE
- MLOps / AIOps
- CI/CD, pipelines
- Deployment patterns
  - Pickle and load, fit()

2. Name the Topics!

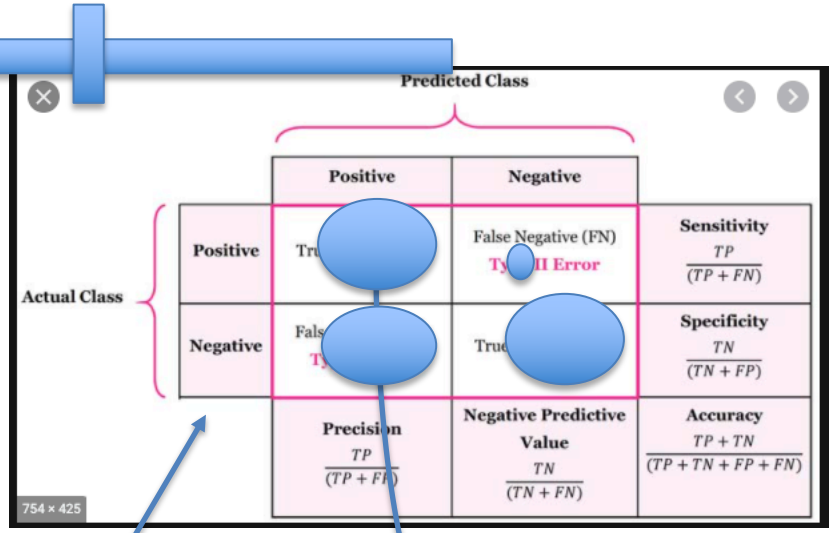
- Topics are clusters
  - Topic 1: world-economy-crisis
- Find the n-gram cluster center, try that as the topic name
- Predict the topic name by training on a dataset
- That dataset has labels that are generated based on the automatic labeling of the cluster by the n-gram that fits the words in the cluster best
- Try with similarity search: https://towardsdatascience.com/how-to-build-a-semantic-search-engine-with-transformers-and-faiss-dcbea307a0e8
- **Automatic Labelling of Topic Models**
- https://stackoverflow.com/questions/32759712/how-to-find-the-closest-word-to-a-vector-using-word2vec
  - Word embedding
    - Word2vec , glove
    - GPT-2

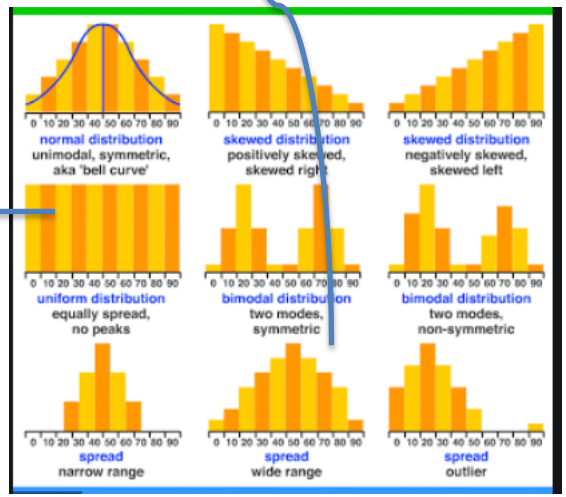


3. Slide out of Confusion!

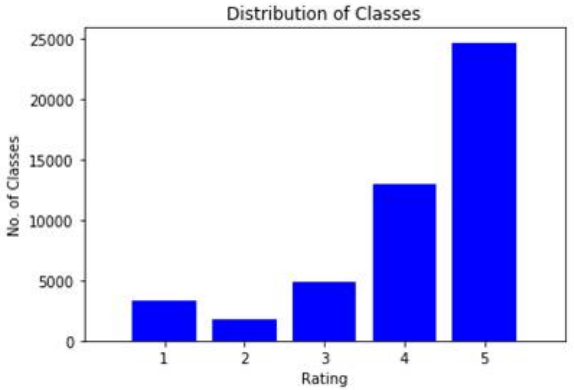
- Create a dashboard for your confusion matrix with sliders
- See code here: https://towardsdatascience.com/data-visualization-with-python-holoviz-plotting-158158e905f2c0

- Data distribution → model.fit() → confusion matrix, AUC/ROC
- What if you changed (sliders on values in the confusion matrix) the values in the confusion matrix?
- What happens to the data columns? Do the distributions on the data have to change (eg 90-10 male-female on credit card dataset becomes 70-30 , increase female data rows ?)
- What happens to the Muller loops?
- Eg, does the model.fit() the data
- Simulate various data (column count , imbalanced data) distributions
- Fit the models
- Plot the confusion matrices
- Compare them to see how the data distribution should be changed if we slide the confusion matrix FP up? What if we want to dec erase FN and increase FP, what should h the data to make that change come about in the cor matrix?
- Is the same algorithm going to give the best f1?





15. Show How Distillation, Amalgamation progressively affect ML Metrics



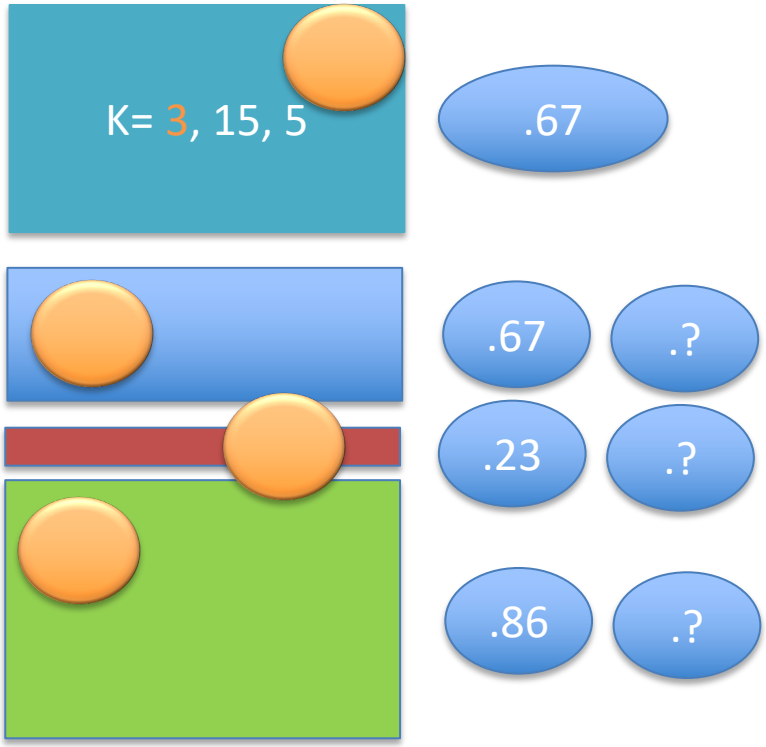
- Given a dataset
- Construct all distillations
- Do three Amalgamations (1 scrape)
- at every step, add a distillation, add an amalgar
- Test muller loop
- Show how the outcomes (confusion matrix, auc, f1, heatmap change)
- Apply class imbalance mitigations, and compare at each stage
- Analyze imbalanced dataset influences on the model metrics: acc, prec, recall, f1, confusion
- https://datascience.stackexchange.com/questions/320156/class-imbalance-problem
- https://medium.com/@dkatzman_3920/class-imbalance-what-is-it-and-how-to-deal-with-it-78768f88a12e

5. Show Impact of Disinformation on Public Health crisis (Pandemic)

- Amalgamation, Distillation
- Topic
- Compare in clusters from various newssources
  - Newsmaxx,fox,cnn|bbc,msnbc, your choice
- Factors in distillation :
  - Contextualization: Nuance: Time-series, focused domain
    - Timeline, location
    - Where did things happen, when
- West coast, East coast, Midwest, South (15 regions, 1 state)
- Choose one of these Angles:
  - Economy:
    - https://towardsdatascience.com/text-ming-comments-of-a-plan-to-get-america-back-to-work-ba82a8153ebe7?source=userActivityShare-c8cbbc37a6fb-160721526715&_branch_match_id=789911520526155835 38
  - Public Health:
    - https://www.sciencedirect.com/science/article/pii/S2590061720300569

6. Compare Cluster Performance

- Cluster your dataset
- Break the dataset into the macros clusters
  - Cluster each macro cluster
  - Compare with diff k's and diff clustering techniques (5 diff, k-means, spectral clustering, gaussian, etc)
- What if you trained your data on different clusters, separately?
- How are each cluster behaving in the confusion matrix, the AUC/ROC, f1, heatmap?
- Compare the performance of various (Muller Loop) models on each cluster
- What is the uncertainty of the prediction for each cluster?
- Conduct Fractal Clustering
- How does the golden cluster change in each of the clusters



DeepContext Data Science Life-cycle ®

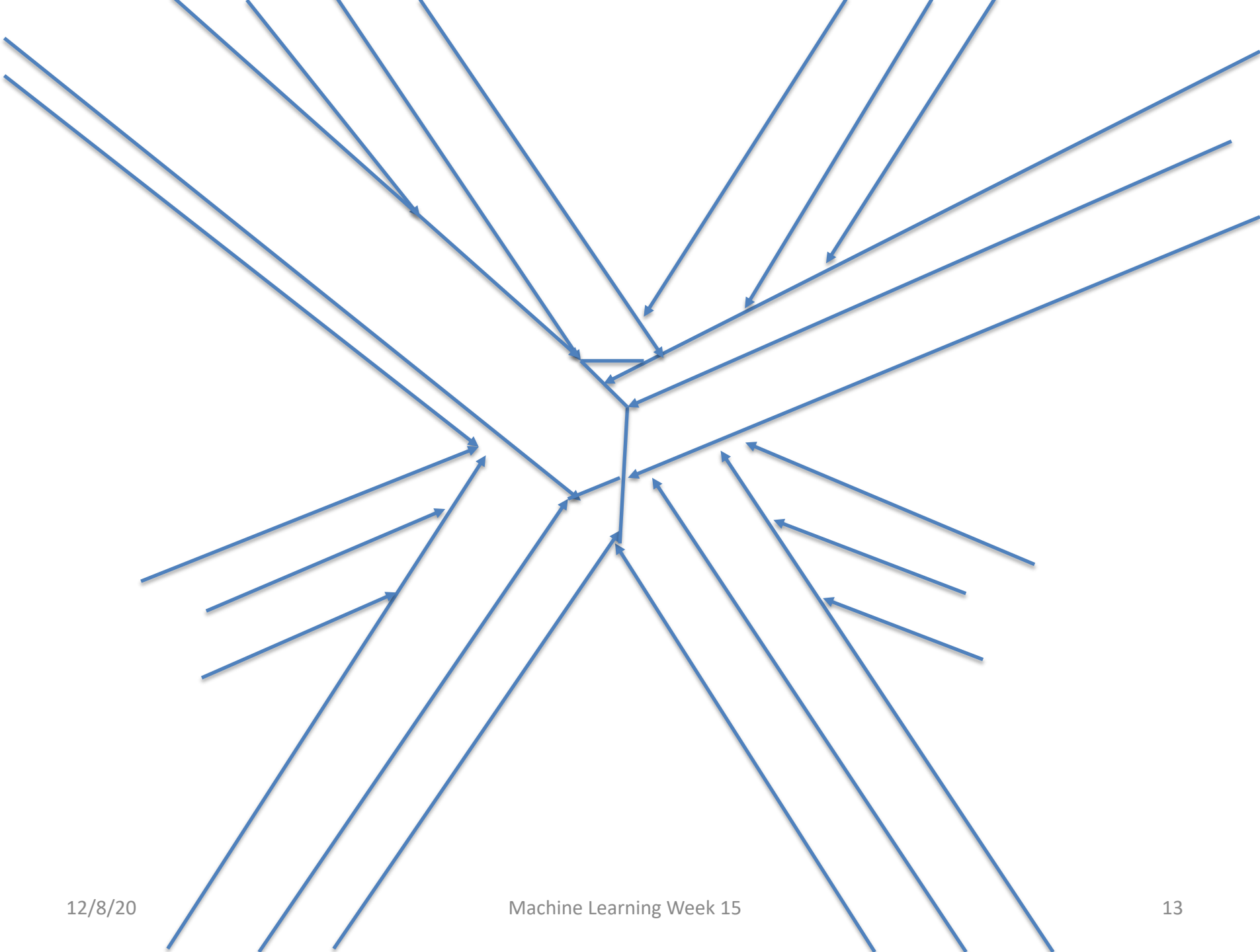
- Data Selection.
  - Find one baseline dataset,
- Data Narrative.
  - Figure out your data narrative
- Algorithm Selection
  - Which algorithms (clustering, regression, classification) are a better fit for your data, and to fulfill your narrative?
- Latent Variables, Models, Manifolds
  - Identify at least one Latent Manifold, which variables should you use? Enrich data in the next step to feed into your latent manifold, to help with explaining results and contributing to your Data Narrative
- Data Enrichment .
  - Find 2 other datasets that will help deepen insights and refine results of regression accuracy
- Amalgamation
  - Apply amalgamation techniques, identify and report back on your amalgamation and how it possibly increased your accuracy, R2, F1 RSME, etc.
- Data Distillation
  - Can you distill your data?
- Curate your data for deeper accuracy, insight
  - Supervised learning, reinforcement learning
- Refine your Data Narrative
  - What should the business do now that you have some insights?
  - What have you discovered?
  - What can you report on?
  - What could you write if you were an investigative reporter or a data detective ?

Data Science Life-cycle ®

- Data Selection.
  - Find one baseline dataset,
- Data Narrative.
  - Figure out your data narrative
- Algorithm Selection
  - Which algorithms (clustering, regression, classification) are a better fit for your data, and to fulfill your narrative?
- Latent Variables, Models, Manifolds
  - Identify at least one Latent Manifold, which variables should you use? Enrich data in the next step to feed into your latent manifold, to help with explaining results and contributing to your Data Narrative
- Data Enrichment .
  - Find 2 other datasets that will help deepen insights and refine results of regression accuracy
- Amalgamation
  - Apply amalgamation techniques, identify and report back on your amalgamation and how it possibly increased your accuracy, R2, F1 RSME, etc.
- Data Distillation
  - Can you distill your data?
- Curate your data for deeper accuracy, insight
  - Supervised learning, reinforcement learning
- Refine your Data Narrative
  - What should the business do now that you have some insights?
  - What have you discovered?
  - What can you report on?
  - What could you write if you were an investigative reporter or a data detective ?

Distillations

1. Customer Identity
2. Entity Resolution
3. Customer Lifetime Value
4. Sentiment
5. Topics
6. Requests
7. Time lines
8. Locations
9. Entities and Relationships Extraction
    - For knowledge graph construction




Deep Context